Q. J. R. Meteorol. Soc. (2001), **127**, pp. 2473–2489

Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size

By DAVID S. RICHARDSON*
European Centre for Medium-Range Weather Forecasts, UK

(Received 14 November 2000; revised 10 May 2001)

SUMMARY

Ensemble forecasts provide probabilistic predictions for the future state of the atmosphere. Usually the probability of a given event E is determined from the fraction of ensemble members which predict the event. Hence there is a degree of sampling error inherent in the predictions. In this paper a theoretical study is made of the effect of ensemble size on forecast performance as measured by a reliability diagram and Brier (skill) score, and on users by using a simple cost–loss decision model. The relationship between skill and value, and a generalized skill score, dependent on the distribution of users, are discussed. The Brier skill score is reduced from its potential level for all finite-sized ensembles. The impact is most significant for small ensembles, especially when the variance of forecast probabilities is also small. The Brier score for a set of deterministic forecasts is a measure of potential predictability, assuming the forecasts are representative selections from a reliable ensemble prediction system (EPS). There is a consistent effect of finite ensemble size on the reliability diagram. Even if the underlying distribution is perfectly reliable, sampling this using only a small number of ensemble members introduces considerable unreliability. There is a consistent over-forecasting which appears as a clockwise tilt of the reliability diagram. It is important to be aware of the expected effect of ensemble size to avoid misinterpreting results. An ensemble of ten or so members should not be expected to provide reliable probability forecasts. Equally, when comparing the performance of different ensemble systems, any difference in ensemble size should be considered before attributing performance differences to other differences between the systems.

The usefulness of an EPS to individual users cannot be deduced from the Brier skill score (nor even directly from the reliability diagram). An EPS with minimal Brier skill may nevertheless be of substantial value to some users, while small differences in skill may hide substantial variation in value. Using a simple cost–loss decision model, the sensitivity of users to differences in ensemble size is shown to depend on the predictability and frequency of the event and on the cost–loss ratio of the user. For an extreme event with low predictability, users with low cost–loss ratio will gain significant benefits from increasing ensemble size from 50 to 100 members, with potential for substantial additional value from further increases in number of members. This sensitivity to large ensemble size is not evident in the Brier skill score. A generalized skill score, dependent on the distribution of users, allows a summary performance measure to be tuned to a particular aspect of EPS performance.

KEYWORDS: Brier score Reliability diagram

1. INTRODUCTION

An ensemble prediction system (EPS) provides a practical method of generating probabilistic forecasts of future weather events. In this paper we consider probability forecasts of a binary weather event E. For example, E could be the occurrence of precipitation higher than a chosen threshold or of temperature less than a given value. For an EPS, the forecast probability of E is usually estimated as the fraction of ensemble members which predict the event.

Because of computational cost, the number of ensemble members may be limited to a few tens at most. The finite size of the ensemble introduces an inevitable degree of sampling error so that the predicted probability will not always be representative of the underlying probability of E.

The performance of a probability forecasting system in predicting E can be evaluated over a large number of cases using a reliability diagram (Wilks 1995). This contains all the information about the forecast and observed probability distributions and is therefore a complete representation of forecast performance (Murphy and Winkler 1987). The Brier (skill) score (Wilks 1995) is a commonly used summary measure of performance.

* Corresponding address: ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK. e-mail: dit@ecmwf.int

In this paper a theoretical study is made of the effect of ensemble size on the Brier score and the reliability diagram. A simple cost–loss decision model is also used to investigate the effect of ensemble size on the economic value of an EPS. By considering the EPS as a random sample taken from an unbiased and representative underlying distribution of possible states, the effect of sampling errors alone on an otherwise perfectly specified EPS can be isolated.

The paper is organized as follows. In section 2, the notion of the EPS as a finite sample drawn from an underlying distribution of forecasts is introduced. The effect of this sampling on the Brier score and reliability diagram are discussed in sections 3 and 4 and the impact on users is considered in section 5. In section 6 the relationship between skill and value is investigated and a generalization of the Brier score is introduced which can be customized to suit different distributions of users. Conclusions are drawn in section 7.

2. Probability forecasts for a finite-sized EPS

Consider probability forecasts of a binary event E. For an EPS, the probability is usually taken as the fraction of ensemble members predicting E. So, for an ensemble of size M, the forecast probabilities P_f will be restricted to a finite set

$$P_f \in \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\right\} = \{p_k, (k = 0, \ldots, M)\}. \tag{1}$$

The EPS can be considered as a random sample of size M taken from an underlying distribution of possible forecasts, with P_f an estimate of the underlying probability P_∞ (the subscript ∞ indicates this probability would be obtained for a sufficiently large ensemble: the limit as $M \to \infty$). Unlike P_f, the underlying probability can vary continuously between 0 and 1. Because of sampling effects, the forecast probability given by the ensemble, P_f, will sometimes be substantially different from P_∞, particularly for small M. This can have a significant effect on EPS verification measures, which may be misinterpreted if the effect of ensemble size is not taken into account.

Over a large number of cases, let g_k be the frequency with which $P_f = p_k$ and let o_k be the observed frequency of E when $P_f = p_k$. The discrete frequencies g_k and o_k depend on the EPS probabilities P_f and are therefore also subject to the effect of sampling. For a given value of the underlying probability $P_\infty = p$, the relative frequency with which k members forecast E is given by the binomial distribution

$$q_k(p) = \Pr\left(P_f = \frac{k}{M} \middle| P_\infty = p\right) = \binom{M}{k} p^k (1-p)^{M-k}, \tag{2}$$

where $\Pr(X|Y)$ stands for probability of X given Y.

Let the distribution of P_∞ be given by the continuous probability density function (p.d.f.) $g(p)$ and let $o(p)$ be the p.d.f. for the conditional probability that E occurs given P_∞. The discrete frequencies g_k and o_k can now be written in terms of the continuous p.d.f.s. $g(p)$ and $o(p)$ as

$$g_k = \int_0^1 q_k(p) g(p)\, \mathrm{d}p \tag{3}$$

and

$$o_k = \left(\frac{1}{g_k}\right) \int_0^1 q_k(p) g(p) o(p)\, \mathrm{d}p. \tag{4}$$

We can use Eqs. (2) and (3) to compare the mean and variance of the EPS and underlying probabilities. The mean of the distribution of P_f is the same as that of P_∞

$$\sum_{k=0}^{M} p_k g_k = \int_0^1 \left(\sum_{k=1}^{M+1} p_k q_k \right) g(p)\, dp = \int_0^1 p g(p)\, dp = \mu \tag{5}$$

(where we have substituted for g_k and noted that the summation is equal to p, from the binomial distribution). However, the variance of the EPS probabilities is larger than that of P_∞

$$\begin{aligned} \left\{ \sum_{k=0}^{M} p_k^2 g_k - \mu^2 \right\} &= \int_0^1 \left(\sum_{k=1}^{M+1} p_k^2 q_k \right) g(p)\, dp - \mu^2 \\ &= \left\{ \int_0^1 p^2 g(p)\, dp - \mu^2 \right\} + (1/M) \int_0^1 p(1-p) g(p)\, dp \end{aligned} \tag{6}$$

(again the summation is known from the mean and variance of the binomial distribution).

The finite ensemble size means that P_∞ is generally misrepresented by P_f. This distorts the frequency with which different probabilities are predicted (g_k), resulting in an increased variance in the EPS probabilities relative to the variance of P_∞. While there is no overall bias in the EPS (E is predicted as often as it occurs in the underlying distribution), sampling errors introduce a tendency for P_f to be overconfident (closer to 0 or 1) compared with P_∞. Because the forecast probabilities are distorted, the observed relative frequencies o_k are also different from those of the underlying distribution given by $o(p)$.

The reliability diagram is a plot of o_k against p_k; g_k is displayed either by labelling the points on the reliability diagram or separately as a histogram. This information is a complete representation of the performance of an EPS in predicting the event E. However, to construct the reliability diagram we need to evaluate the integrals in Eqs. (3) and (4) for which we must specify $g(p)$. Before we do this (in section 4) we first consider the overall skill of the EPS as measured by the Brier score.

3. Ensemble size and the Brier score

The Brier score for an M-member EPS can be written as

$$b_M = \sum_{k=0}^{M} (p_k - 1)^2 g_k o_k + \sum_{k=0}^{M} p_k^2 g_k (1 - o_k). \tag{7}$$

Similarly, the Brier score b_∞ for the continuously distributed probability P_∞ is

$$b_\infty = \int_0^1 (p-1)^2 g(p) o(p)\, dp + \int_0^1 p^2 g(p)(1 - o(p))\, dp. \tag{8}$$

By substituting for g_k and o_k into Eq. (7), we find that

$$b_M = b_\infty + \frac{1}{M} \int_0^1 p(1-p) g(p)\, dp. \tag{9}$$

The integral in Eq. (9) is positive definite, so the finiteness of the ensemble always has an adverse effect on the Brier score. The second term in Eq. (9) is exactly the increase in variance of the EPS probabilities noted in section 2 (Eq. (6)).

The Brier skill score relative to climatology is defined as

$$B = 1 - \frac{b}{\bar{o}(1-\bar{o})}, \tag{10}$$

where $\bar{o}$ is the (sample) climatological frequency of E

$$\bar{o} = \sum_{k=0}^{M} o_k g_k = \int_0^1 o(p)g(p)\,\mathrm{d}p. \tag{11}$$

$B = 0$ for climatological forecasts ($P_f \equiv \bar{o}$), and $B > 0$ indicates a forecasting system has skill relative to climatology. Maximum skill, $B = 1$, is obtained for perfect deterministic forecasts.

The Brier skill score for the EPS is

$$\begin{aligned} B_M &= B_\infty - \frac{1}{M\bar{o}(1-\bar{o})}\int_0^1 p(1-p)g(p)\,\mathrm{d}p \\ &= B_\infty - \frac{1}{M}\left(\frac{\mu(1-\mu)}{\bar{o}(1-\bar{o})} - \frac{\sigma^2}{\bar{o}(1-\bar{o})}\right) \end{aligned} \tag{12}$$

where μ and σ^2 are the mean and variance of the distribution of P_∞.

The magnitude of the effect on B_M depends on both the ensemble size M and the distribution $g(p)$ of forecast probabilities. The reduction in skill is largest for small M, as expected since small ensembles will have the largest sampling errors. However, Eq. (12) shows that the effect of sampling on the Brier score does not just depend on the size of the ensemble. Skill is also affected by the variability of the forecast probabilities and by how well the underlying forecast distribution represents the 'true' level of predictability of the event.

In the remainder of this paper we will focus solely on the effect of ensemble size on EPS performance and will therefore assume that the distribution of P_∞ may be taken as truth. This assumption will allow us to quantify the effect of ensemble size on the performance of an otherwise perfectly formulated EPS.

(a) *Brier score for a representative EPS*

We assume that the underlying forecast probabilities are perfectly reliable, so that $o(p) = p$ and $\mu = \bar{o}$. The Brier score b_∞ becomes (putting $o(p) = p$ in Eq. (8) and rearranging)

$$b_\infty = -\int_0^1 (p-\bar{o})^2 g(p)\,\mathrm{d}p + \bar{o}(1-\bar{o}) = -\sigma^2 + \bar{o}(1-\bar{o}) \tag{13}$$

and the Brier skill score is

$$B_\infty = \frac{\sigma^2}{\bar{o}(1-\bar{o})}. \tag{14}$$

So B_∞ is the variance of P_∞ normalized by the variance of the observations (uncertainty). An increase in variance indicates a move of probability away from the climatological frequency $\bar{o}$ towards the deterministic extremes of 0 and 1. Given perfect reliability, this necessarily means an increase in predictability (without the assumption of perfect reliability there is no such guarantee).

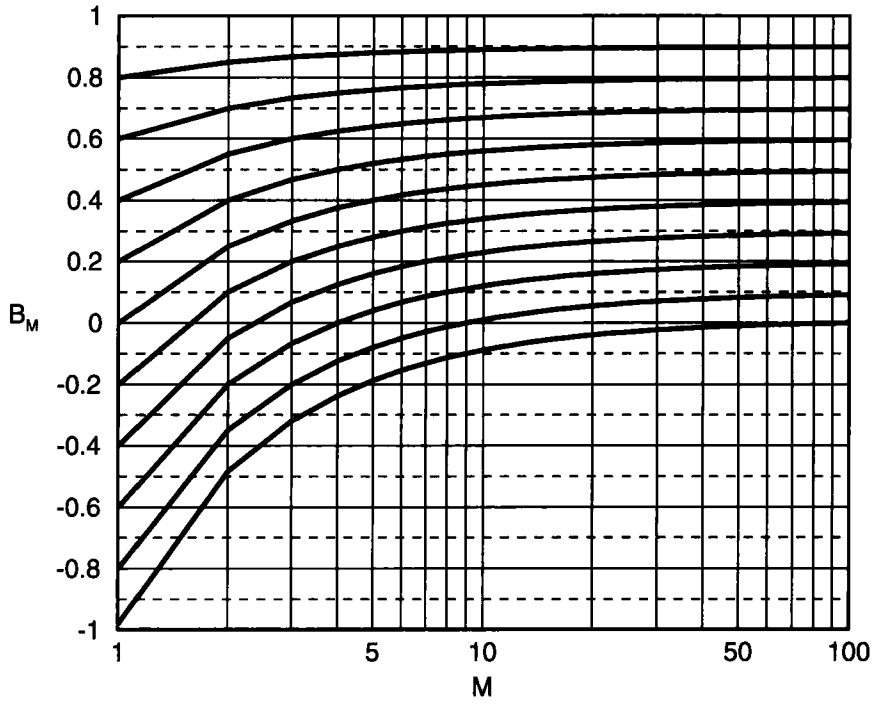

Figure 1. Brier skill score, B_M, as a function of ensemble size for a range of predictabilities. The ten curves show how B_M varies with the number of ensemble members, M, for $B_\infty = 0.01, 0.1, 0.2, 0.3, \ldots, 0.9$ (see text).

From Eq. (12) and (14) we can now write the Brier skill score for an M-member ensemble, B_M, as a function of M and B_∞

$$B_M = B_\infty - \frac{1 - B_\infty}{M} = \frac{(M+1)B_\infty - 1}{M}. \tag{15}$$

The variation of B_M with M is shown in Fig. 1. For all B_∞, B_M increases most rapidly for $M < 10$, and is close to its asymptotic limit for $M \sim 100$. The dependence on $g(p)$ (i.e. σ^2 or B_∞) can be seen. For high predictability (high B_∞), even a single forecast will score well, but for less predictable events larger ensembles are necessary. If predictability is inherently low, then even a perfectly formulated EPS will not perform well unless it has many members. For example, if $B_\infty = 0.1$, then a 50-member ensemble would score about 0.09 while a 10-member ensemble would only just achieve positive skill, and smaller ensembles would be less skilful than climatology.

Although B_∞ is always positive, inevitable sampling errors can result in minimal or negative ensemble skill in some situations. $B_M \leq 0$ does not necessarily mean that there is no predictability; it may just be that a larger ensemble is needed to realize the (probably low) underlying predictability. Under the assumption of reliability of P_∞, Eq. (15) or Fig. 1 can be used to estimate the potential level of skill which could be achieved with a larger ensemble. An empirical example of the variation of B_M with ensemble size is given by Talagrand *et al.* (1997); this agrees well with the theoretical results of Fig. 1. Talagrand *et al.* also briefly discuss the theoretical relationship of Eq. (9).

(*b*) *Brier score for a single deterministic forecast*

An estimate of potential predictability using Eq. (15) can be made even for a set of deterministic forecasts (single-member ensembles). Assuming that the deterministic forecasts are representative elements of a reliable EPS, the expected skill of the EPS is $B_\infty = (1 + B_1)/2$. If, further, the distribution of forecast probabilities is represented by a parametric distribution as in the following section, then it is possible to evaluate all aspects of the potential EPS performance in predicting E from the deterministic forecasts.

In principle, this is a possible alternative to the 'perfect EPS' estimate of potential predictability, where a set of ensemble forecasts is verified against a randomly chosen ensemble member rather than analyses. This alternative approach would allow either a much larger sample of cases for the same computational cost, or estimates to be made on the same number of cases at a much reduced cost. The results would indicate potential levels of performance. From this the ensemble size needed for any required levels of skill could be deduced. Practical application of the results would depend on how well an EPS based on the deterministic forecasts satisfies the assumptions of representativeness and reliability.

4. Ensemble size and the reliability diagram

To examine the effect of ensemble size on the reliability diagram, we need to evaluate g_k and o_k as functions of $g(p)$ (Eqs. (3) and (4)). To do this we specify the p.d.f. $g(p)$ using a beta distribution, the standard choice to represent a probability variable (Wilks 1995; Wilks and Hamill 1995):

$$g(p) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} p^{r-1}(1-p)^{s-1}, \tag{16}$$

where r and s are the two parameters which determine the shape of the distribution and Γ is the gamma function (Wilks 1995).

The mean and variance of this distribution are, respectively, given by

$$\mu = \frac{r}{r+s} \tag{17}$$

and

$$\sigma^2 = \frac{rs}{(r+s)^2(r+s+1)}. \tag{18}$$

Again, we assume perfect reliability so that μ is equal to the observed climate frequency $\bar{o}$. For a particular event, $\bar{o}$ is fixed. This then specifies the mean for the beta distribution which in turn fixes the ratio between r and s. The p.d.f. $g(p)$ then varies with σ^2, larger σ^2 indicating higher predictability.

Figure 2 shows four examples of $g(p)$, all with the same mean $\bar{o} = 0.2$. Complete unpredictability ($\sigma^2 = 0$) would be represented by a single (infinite) spike at $p = \bar{o}$. As σ^2 increases, the distribution spreads out from the climate probability. For small σ^2 the distribution is approximately normal; as σ^2 increases the distribution becomes skewed towards $p = 0$. When $r = 1$, the distribution becomes a decreasing function of p: many high-confidence forecasts are made that E will not occur, but there are few forecasts with high probability that E will occur. If both r and s are less than one, $g(p)$ is U-shaped with the probability concentrated towards the extremes of 0 and 1, representing high predictability.

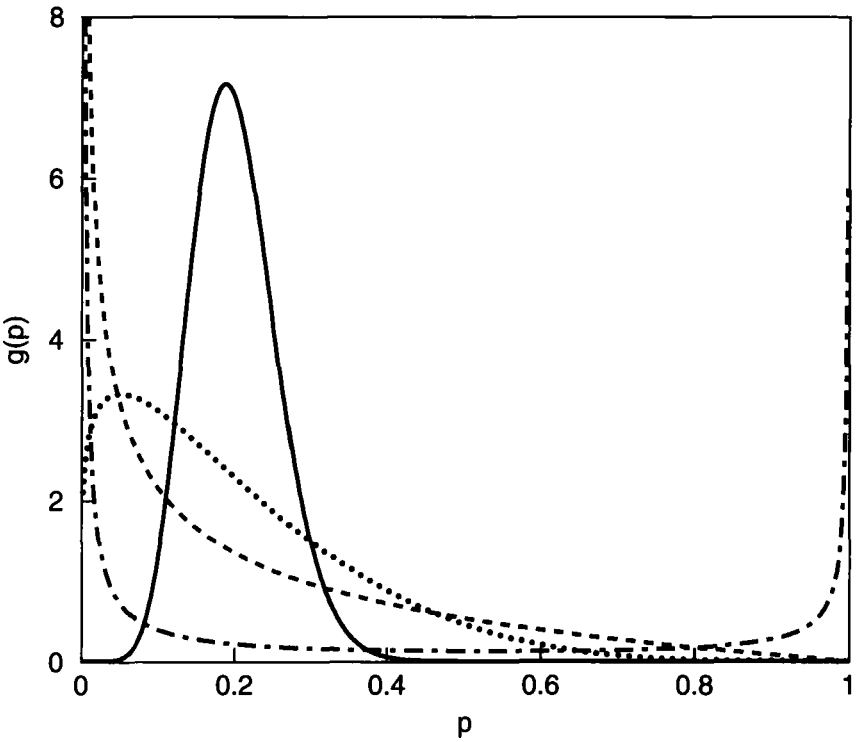

Figure 2. Examples of the beta distribution for four different choices of the distribution parameters r and s (see text). The mean is the same for all cases ($\mu = 0.2$). Full line: $r = 10$, $s = 40$; dotted line: $r = 1.2$, $s = 4.8$; dashed line: $r = 0.5$, $s = 2.0$; chain-dashed line: $r = 0.05$, $s = 0.2$. See text for explanation of axes.

For a finite ensemble, g_k and o_k can now be written in terms of the parameters of $g(p)$ (details are given in an appendix). The difference between the discrete forecast probabilities p_k and the corresponding observed relative frequencies o_k is

$$o_k - p_k = \left(\frac{1 - B_\infty}{1 + (M-1)B_\infty} \right) (\bar{o} - p_k). \tag{19}$$

Equation (19) quantifies the effect of ensemble size on the reliability diagram. Although we have defined the underlying forecast probabilities P_∞ to be perfectly reliable, the ensemble probabilities will be increasingly unreliable as they get further away from $\bar{o}$. For lower probability, o_k is greater than p_k, and vice versa for larger probability. This is seen in the reliability diagram as a clockwise tilt away from the diagonal. It is interpreted as a tendency for the forecasts to be 'overconfident' in that they too often predict E with more certainty than is warranted.

The effect is illustrated using two example p.d.f.s for P_∞, representing two different events with the same underlying level of predictability. Setting ($r = s = 3$) gives $\bar{o} = 0.5$ and $B_\infty = 1/7 \approx 0.14$, while ($r = 1.2$, $s = 4.8$) gives $\bar{o} = 0.2$ and again $B_\infty = 1/7$. Reliability diagrams for 10- and 50-member ensembles are shown in Figs. 3 and 4. In all cases the curves of o_k deviate, often substantially, from the diagonal line of perfect reliability. This appears as a rotation about the climatological frequency, showing that the forecasts are consistently overconfident. Most forecast probabilities cannot, therefore, be taken at face value, despite the imposition of perfect reliability for the underlying continuous probabilities P_∞. The distance from the diagonal increases as p_k

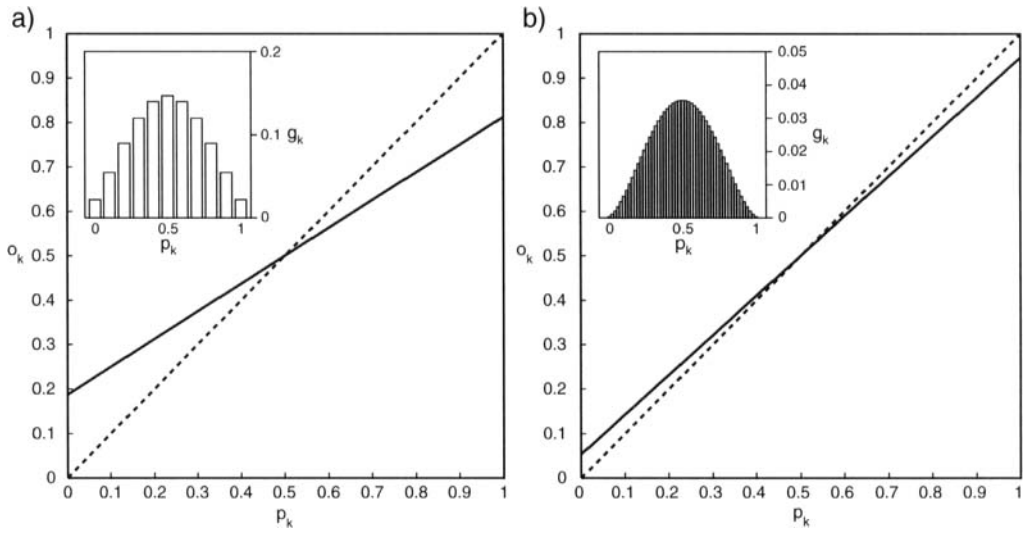

Figure 3. Reliability diagrams for theoretical ensemble forecasts for (a) a 10-member ensemble prediction system (EPS) and (b) a 50-member EPS. Distribution of underlying forecast probabilities is completely reliable and specified by a beta distribution with $r = s = 3$. See text for details and explanation of symbols.

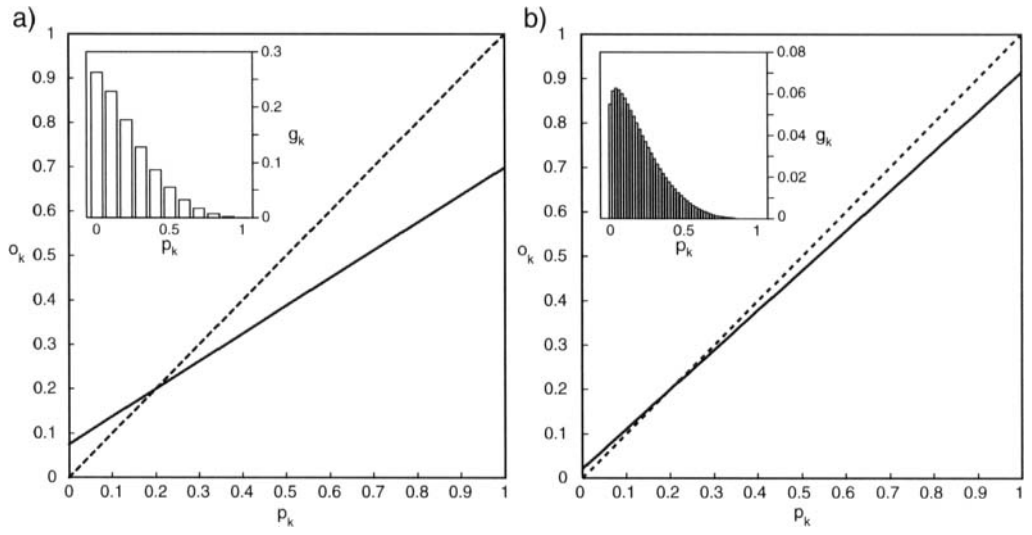

Figure 4. As Fig. 3 but with $r = 1.2$ and $s = 4.8$.

becomes further from $\bar{o}$, so the overconfidence is greatest for high-confidence forecasts of the less common event.

The curves for the 50-member EPS are generally much closer to the diagonal than those for the 10-member EPS, demonstrating the improved reliability resulting from the better sampling of the larger ensemble. This improved reliability is reflected in the higher skill for both events: B is about 0.13 for 50 members, but less than 0.06 for 10 members. Although the skill of the 50-member EPS is close to its asymptotic limit $B_\infty = 0.14$, the noticeable overconfidence for the more extreme probability categories

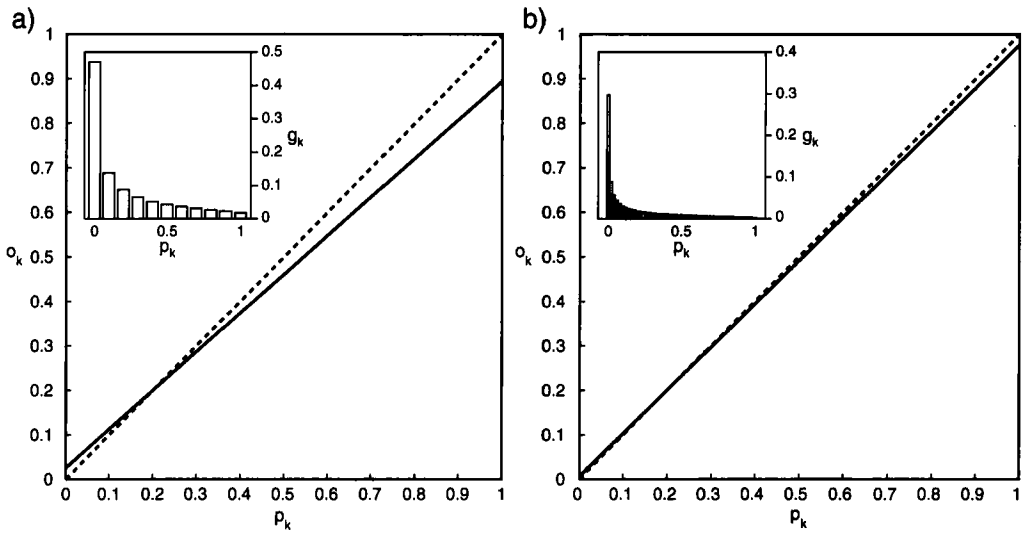

Figure 5. As Fig. 3 but with $r = 0.3$ and $s = 1.2$.

may be of considerable significance to some users. While the scores and reliability diagrams indicate better performance for the 50-member EPS, the histograms of g_k for both events are sharper and closer to $\overline{o}$ than for the 10-member EPS. In Fig. 4, for example, $p = 0$ is the most frequent 10-member EPS probability, while for 50 members the peak is shifted closer to the position in the underlying p.d.f. (cf. Fig. 2). Figure 5 shows reliability diagrams for EPS forecasts of a more predictable event, with $B_\infty = 0.4$ ($r = 0.3$, $s = 1.2$; $\overline{o} = 0.2$). The 50-member EPS is now fairly reliable, but there is still significant overconfidence for a 10-member EPS.

5. Ensemble size and potential economic value

We have seen that sampling errors inevitably introduce a degree of unreliability into any finite-sized EPS. While it is clear that a 10-member EPS is affected more than a 50-member EPS, the impact on the usefulness of the forecasts is not obvious. For example, are the relatively small errors of the 50-member EPS in Fig. 4 acceptable for users or will there be significant benefit to be obtained from increasing ensemble size further?

To address the potential benefit of an EPS to different users we use the value diagnostic, derived from a simple cost–loss model of economic decision making (Murphy 1977; Richardson 2000). Consider a decision maker who will incur a loss L if the event E occurs and no protective action has been taken but who has the option of taking protective action at a cost C to prevent this potential loss. If no forecasts were available, the decision maker would either always or never protect, whichever gives the lowest expected loss. The decision maker uses forecasts to decide whether or not to take protective action on each occasion. If the forecasts are expressed as probabilities, the decision to act is taken once the forecast probability exceeds a certain threshold.

The value V of an EPS is defined as the savings made by using the EPS as a fraction of the potential savings which would be achieved with perfect forecast information. $V = 0$ indicates that forecasts have no more value that can be achieved from knowing only $\overline{o}$. V depends not only on the performance of the forecast system, but also on $\overline{o}$ and

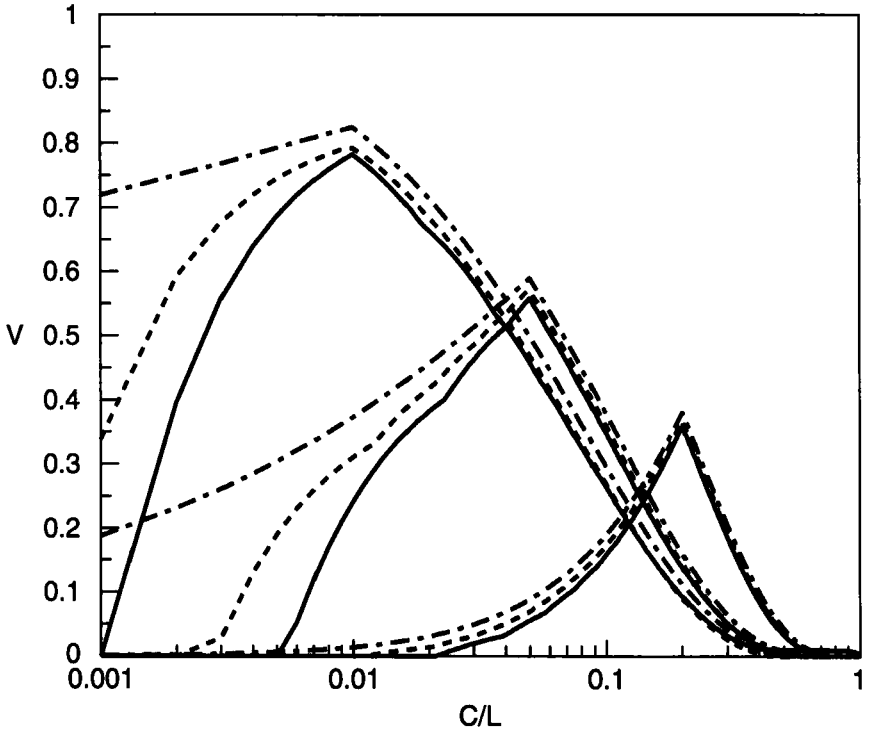

Figure 6. Potential value V as a function of cost–loss ratio C/L for three events: $\bar{o} = 0.01$ (left-most set of curves), $\bar{o} = 0.05$ (centre), and $\bar{o} = 0.2$ (right-most curves) (see text). The Brier skill score for the underlying forecasts is the same for each event, $B_{\infty} = 1/7$. The three curves in each set are for a 50-member ensemble prediction system (EPS) (full line), 100-member EPS (dashed line), and the potential value of the underlying distribution (limit for large ensemble, dash-dotted line).

on the user's cost–loss ratio $\alpha = C/L$: optimal value will be achieved if action is taken whenever the probability of E occurring is greater than α (Richardson 2000). For the finite-sized ensembles this optimal probability is o_k rather than p_k, in effect calibrating the forecasts to compensate for sampling errors.

As in the previous section, we use a beta distribution to specify $g(p)$ and assume the underlying distribution is reliable. The maximum potential value can be calculated directly from the parameters of $g(p)$, while the value of an M-member EPS is calculated from g_k and o_k.

Figure 6 shows V for a number of ensemble systems with $B_{\infty} = 1/7$. To show the effect of the observed frequency of the event, results are shown for three events with $\bar{o} = 0.2$ (cf. Fig. 4), 0.05 and 0.01. For each event the value of a 50-member EPS is compared with that of a 100-member EPS and with the potential value obtainable with a large-enough EPS. There is some potential for improvement to be gained by increasing ensemble size for all events, but the benefit is substantially larger for the rarer events. It is worth contrasting the small differences in skill for the different ensemble sizes with the sometimes substantial impact on the user. B_{50} is less than 2% smaller than B_{∞} (0.126 compared with 0.143). For some users the difference in value is also small, but for others the 50-member EPS has no value while the potential from a larger ensemble is more than 70%. (Fig. 6).

The potential usefulness of an EPS with very small Brier skill is shown in Fig. 7. The same three events as shown in Fig. 6 are considered, but this time the skill of the

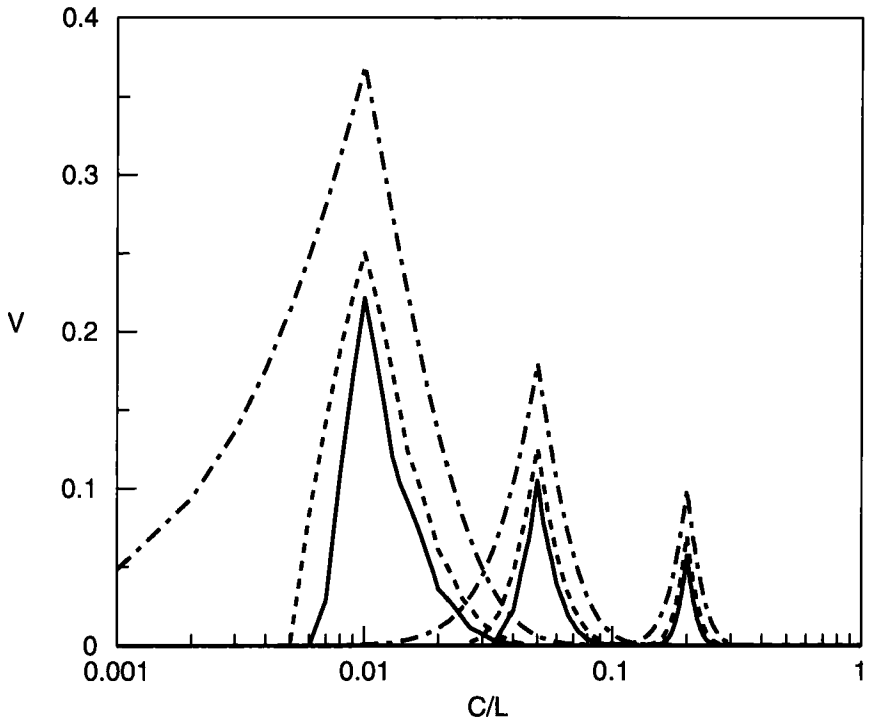

Figure 7. As Fig. 6 but for $B_\infty = 0.01$ (see text).

underlying distribution is chosen to be close to zero, $B_\infty = 0.01$. With this potential level of skill, the scores for the 50- and 100-member ensembles are $B_{50} = -0.01$ and $B_{100} = 0.0$. Despite the absence of overall skill for either EPS, both systems have positive value for a range of users. Again, it can be seen that increasing ensemble size would give significant benefits, particularly to users with low cost–loss ratios for extreme events.

6. THE RELATIONSHIP BETWEEN SKILL AND VALUE

The value curves of Figs. 6 and 7 show that the benefit of an EPS can vary greatly between different users. Therefore there can be no simple relationship between the Brier skill score (a single overall measure) and the value to specific users. On the other hand, B can be interpreted as a measure of overall value if we assume that the cost–loss ratios (α) of users are distributed uniformly on the interval (0,1) (Murphy 1966).

The true distribution of users is not well known, although there are indications that it is unlikely to be uniform and will more probably have larger weight at small α (Roebber and Bosart 1996). Given that value and sensitivity to ensemble size vary greatly over the full range of α, it is important to consider the implications of the assumption of uniformity implicit in the Brier score. In this section we define a measure of overall value based on an arbitrary distribution of users and show it can be expressed in terms corresponding to reliability and resolution. We show how the generalized skill score varies depending on the choice of user distribution.

(a) *A general measure of overall skill or value*

Assume the distribution of users is given by a p.d.f. $u(\alpha)$. We can then derive a measure of skill or value based on the total saving made by all users. For this general derivation we assume that a user will take the EPS probabilities at face value so that each user will take action when p is greater than their cost–loss ratio α. For an M-member EPS, the total expense over all users can be written as

$$T_F = T_C + \sum_{k=0}^{M} g_k \int_{o_k}^{p_k} u(\alpha)(\alpha - o_k)\, d\alpha - \sum_{k=0}^{M} g_k \int_{\bar{o}}^{o_k} u(\alpha)(o_k - \alpha)\, d\alpha \qquad (20)$$

where T_C is the total expense if all users act using only the climatological probability $\bar{o}$. The second and third terms are general forms of reliability and resolution components of the Brier score (Wilks 1995). Details of the derivation and the relationship to the Brier score are given in an appendix.

The last term in Eq. (20) is the maximum reduction in expense which would be achieved if all users act when the actual probability of the event, o_k, is greater than the relevant cost–loss ratio α. This benefit increases as the probabilities become further away from the climatological frequency (note $\bar{o}$ as a limit in the integral; cf. resolution). The potential benefit is reduced if users act on forecast probabilities which are not completely reliable. The second term in Eq. (20) indicates the additional expense incurred. This reliability term depends on the difference between p_k and o_k (the limits in the integral) and decreases as this difference reduces since there are then fewer occasions on which the incorrect choice of action is made.

(b) *Generalized Brier score and ensemble size for a representative EPS*

In section 3(a) it was shown that if an EPS is considered as a representative sample from a reliable forecast p.d.f., B_M increases monotonically with ensemble size (as M increases, resolution grows while the reliability term decreases). B_M can increase substantially with M for small ensembles, but there is relatively little change once ensemble size increases beyond about 50 members.

We now consider the corresponding situation for the generalized skill score, G_M, for a number of different distributions of users. To illustrate, we consider distributions of users given by the four sample beta distributions of Fig. 2. The variation of G_M with ensemble size for these user distributions is shown in Fig. 8, for a rare event with $\bar{o} = 0.01$ (unlike B, the generalized skill also depends on the observed frequency of the event).

While resolution generally increases with both M and B_∞, the behaviour of the reliability term is more difficult to predict. This term can increase for some users as M increases and result in a reduction in skill for a larger ensemble (Figs. 8(a) and (b)). It is also possible that for a fixed ensemble size, G_M will decrease as B_∞ increases, again due to poorer reliability for certain users (there is not an example of this in Fig. 8, although Fig. 8(a) gives an indication of the possibility). Such 'skill–value reversals' have been noted by Murphy and Ehrendorfer (1987) for an imperfect single-member EPS. In our case, the assumption of reliability of the underlying probabilities precludes these skill–value reversals for $M = 1$, but does allow this possibility for larger M. If the majority of users have relatively small α, there may be substantial benefits to be gained from very large ensembles; G_M will reflect this sensitivity (Fig. 8(d)).

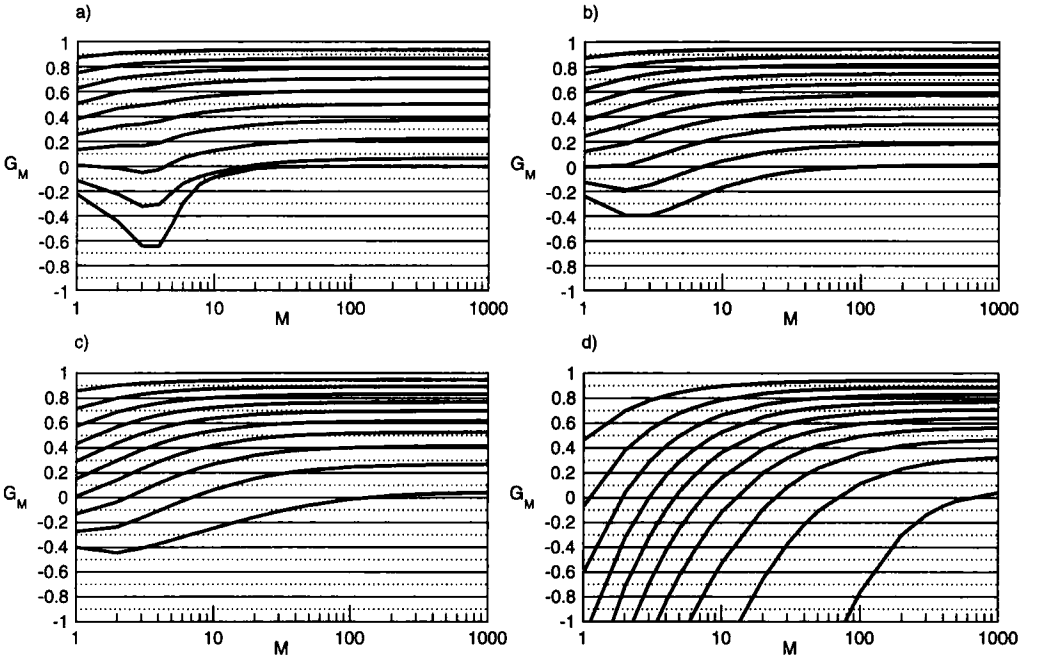

Figure 8. The variation of a generalized skill score G_M with ensemble size M for four different distributions of users, specified by the four beta functions shown in Fig. 2. The ten curves in each panel correspond to the curves for B_M in Fig. 1. Note that the horizontal scale has been extended to allow for the greater sensitivity to ensemble size.

7. Conclusions

We have studied the effect of ensemble size on probability forecasts of binary events. For an EPS, the probability of a given event E is usually taken as the fraction of ensemble members predicting E. Hence there is a degree of sampling error inherent in the EPS probabilities. This sampling error distorts the frequency with which different probabilities are predicted and introduces an inevitable overconfidence in the forecasts.

The Brier skill score, B_M, for any finite-sized EPS will be below its potential value (large-ensemble limit). The effect depends on both the ensemble size and the distribution of forecast probabilities. To investigate the effect of ensemble size alone, we considered ensembles as samples from perfectly reliable underlying distributions. With this assumption, B_M depends only on M and B_∞, where B_∞ itself can be interpreted as the normalized variance of forecast probabilities, a measurement of predictability. For low-predictability events, B_M is unlikely to be positive for $M < 10$.

To quantify the effect of ensemble size on the reliability diagram, the distribution of forecast probabilities was specified using a beta distribution. There are characteristics to the reliability diagram common to all finite-sized ensembles. There will always be a clockwise tilt away from the diagonal indicating that the EPS is overconfident, while the distribution of EPS probabilities (g_k) will be wider than for the underlying distribution. It is tempting to interpret overconfidence on a reliability diagram as an indication of too little ensemble spread. However, while small spread would lead to overconfidence, the inevitable effect of sampling error must be taken into account before an EPS can be considered to be inherently lacking spread. A small ensemble should not be expected to provide reliable forecasts.

As a consequence of the intrinsic unreliability of a finite-sized ensemble, the direct use of ensemble probabilities will be sub-optimal. To maximize the benefit to end users, forecast probabilities will need to be adjusted to compensate for this unreliability. The usefulness of an EPS to individual customers cannot be deduced from the Brier skill score (nor even directly from the reliability diagram). A simple cost–loss model decision model was used to examine the impact of finite ensemble size on users. Optimal value is achieved by using the observed relative frequency, o_k, rather than the forecast probability, p_k, to define the decision threshold. This is effectively a calibration of the forecast probabilities to compensate for the intrinsic unreliability.

The sensitivity of users to differences in ensemble size depends on the predictability and frequency of the event and on the cost–loss ratio of the user. An EPS with $B < 0$ may nevertheless be of substantial value to some users, while small differences in skill may hide substantial variation in value. For example, although there is little change in B once an EPS has 50 or so members, for an extreme event with low predictability, low C/L users will gain significant benefits from increasing ensemble size from 50 to 100 members, with potential for substantial additional value from further increases in number of members.

The relationship between value and skill is inevitably complex. The overall value of an EPS depends on the distribution of users. For a given distribution of users, the total savings relative to climatology can be used to define a measure of skill. Changes in skill are then direct measures of the change in overall benefit to the given set of users. It should be noted that this does not guarantee that all users will benefit equally; it is possible that some users will be worse off even though the overall change is positive. The Brier skill score is a special case of this overall value where the distribution of users is assumed to be uniform throughout the interval (0,1).

The generalized score can be written in terms which may be interpreted as user-dependent measures of reliability and resolution. Resolution is the saving which would be made if each user acted at the optimal probability threshold (given by o_k). The reliability term defines the additional expense incurred when users accept the forecast probabilities at face value (and therefore act when $\alpha > p_k$). While resolution generally increases with both M and B_∞, the behaviour of the reliability term is more difficult to predict. The variation of the generalized user-dependent skill score G_M with ensemble size can be substantially different from that of B_M.

The suitability of a particular performance measure depends on the aspects of performance under investigation. For example, B_M provides an estimate of the underlying predictability of an event, but it is not an appropriate measure to evaluate the potential benefits of a very large ensemble. Large ensembles will be of potential use to a particular group of decision makers (for example those with small cost–loss ratios). For such users there may be significant benefits to be obtained from ensembles with several hundred members. To measure the usefulness of such an EPS, a more specific evaluation measure is needed such as the value diagnostic or the generalized skill score G with an appropriate choice of user distribution.

ACKNOWLEDGEMENTS

I thank Tim Palmer and members of the predictability group at the ECMWF for numerous helpful discussions during the course of this work.

APPENDIX A

For an M-member ensemble we evaluate g_k, the frequency of occurrence of each probability class, and the corresponding observed relative frequency o_k. Substituting from Eqs. (2) and (16) into Eq. (3) we obtain

$$g_k = \binom{M}{k} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^1 p^k (1-p)^{M-k} p^{r-1} (1-p)^{s-1} \, dp. \tag{A.1}$$

Noting that the integral is another beta function $B(r+k, s+M-k)$ we find

$$g_k = \binom{M}{k} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{\Gamma(r+k)\Gamma(s+M-k)}{\Gamma(r+s+M)}. \tag{A.2}$$

We can similarly deduce an expression for o_k, and using g_k from above we find that

$$o_k = \frac{\Gamma(r+k+1)\Gamma(r+s+M)}{\Gamma(r+k)\Gamma(r+s+M+1)} = \frac{r+k}{r+s+M}. \tag{A.3}$$

The difference between the forecast probability $p_k = k/M$ and the observed relative frequency o_k can now be written as

$$o_k - p_k = \frac{r-(r+s)p_k}{r+s+M} = \frac{\mu - p_k}{1+\{M/(r+s)\}}. \tag{A.4}$$

Noting that the Brier skill score for the forecast probabilities can be written in terms of the beta distribution parameters as

$$B_\infty = \frac{1}{r+s+1} \tag{A.5}$$

this can be rewritten as

$$o_k - p_k = \frac{\bar{o} - p_k}{1 + (MB_\infty)/(1-B_\infty)} = \left(\frac{1-B_\infty}{1+(M-1)B_\infty} \right) (\bar{o} - p_k). \tag{A.6}$$

APPENDIX B

A general measure of overall value

Assume the distribution of users is given by a p.d.f. $u(\alpha)$. We will derive a measure of skill or value based on the total saving made by all users. For this general derivation we assume that a user will take the EPS probabilities at face value so that each user will take action when the forecast probability p is greater than their cost–loss ratio α.

Consider the occasions when the forecast probability is p. Users with cost–loss ratio α less than p will take action and hence incur cost α (per unit loss). The total cost over all these users is

$$X = \int_0^p u(\alpha)\alpha \, d\alpha. \tag{B.1}$$

All other users will not act and will incur loss ($L = 1$) when the event occurs. The total loss for these users is

$$X = o(p) \int_p^1 u(\alpha) \, d\alpha = \bar{o} - o(p) \int_0^p u(\alpha) \, d\alpha. \tag{B.2}$$

The overall expense over all users and all forecast probabilities is then

$$T_F = \int_0^1 g(p) \int_0^p u(\alpha)(\alpha - o(p))\, d\alpha\, dp + \bar{o}. \tag{B.3}$$

We now split the inner integral into three parts:

$$\begin{aligned} T_F = \int_0^1 g(p) \Bigg\{ & \int_0^{\bar{o}} u(\alpha)(\alpha - o(p))\, d\alpha\, dp + \int_{\bar{o}}^{o(p)} u(\alpha)(\alpha - o(p))\, d\alpha\, dp \\ & + \int_{o(p)}^{p} u(\alpha)(\alpha - o(p))\, d\alpha\, dp \Bigg\} + \bar{o}. \end{aligned} \tag{B.4}$$

The first term is related to the climate expense T_C:

$$\begin{aligned} \int_0^1 g(p) \int_0^{\bar{o}} u(\alpha)(\alpha - o(p))\, d\alpha\, dp &= \left(\int_0^{\bar{o}} u(\alpha)\alpha\, d\alpha \right) - \bar{o}\left(1 - \int_{\bar{o}}^1 u(\alpha)\, d\alpha \right) \\ &= T_C - \bar{o} \end{aligned} \tag{B.5}$$

so that

$$\begin{aligned} T_F = T_C &+ \int_0^1 g(p) \int_{o(p)}^p u(\alpha)(\alpha - o(p))\, d\alpha\, dp \\ &- \int_0^1 g(p) \int_{\bar{o}}^{o(p)} u(\alpha)(o(p) - \alpha)\, d\alpha\, dp. \end{aligned} \tag{B.6}$$

For perfect forecasts, the total expense over all users is

$$T_P = \int_0^1 u(\alpha)\alpha\bar{o}\, d\alpha = \bar{o}\,\bar{\alpha}. \tag{B.7}$$

A measure of overall value can then be defined as

$$G = \frac{T_C - T_F}{T_C - T_P}. \tag{B.8}$$

For a finite-sized EPS, the equivalent of Eq. (B.6) is

$$T_F = \sum_{k=0}^{M} g_k \int_{o_k}^{p_k} u(\alpha)(\alpha - o_k)\, d\alpha - \sum_{k=0}^{M} g_k \int_{\bar{o}}^{o_k} u(\alpha)(o_k - \alpha)\, d\alpha + T_C. \tag{B.9}$$

To see the relationship to the Brier score, we choose the distribution of users $u(\alpha)$ to be uniform over (0,1). Eq. (B.9) then becomes

$$\begin{aligned} T_F &= \sum_{k=0}^{M} g_k \frac{1}{2}(p_k - o_k)^2 - \sum_{k=0}^{M} g_k \frac{1}{2}(o_k - \bar{o})^2 + \frac{1}{2}\bar{o}(1 - \bar{o}) + \frac{1}{2}\bar{o} \\ &= \frac{1}{2}(b_{rel} - b_{res} + b_{unc}) + \frac{1}{2}\bar{o} \end{aligned} \tag{B.10}$$

where the bracketed term on the right-hand side is the Brier score expressed in the standard decomposition in terms of reliability, resolution and uncertainty (Wilks 1995). It is easily seen that for this distribution of users, $G = B$ so that the Brier skill score is the overall value for a uniformly distributed set of users.

REFERENCES

Murphy, A. H.	1966	A note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation. *J. Appl. Meteorol.*, **5,** 534–537
	1977	The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Weather Rev.*, **105,** 803–816
Murphy, A. H. and Ehrendorfer, M.	1987	On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Weather and Forecasting*, **2,** 243–251
Murphy, A. H. and Winkler, R. L.	1987	A general framework for forecast verification. *Mon. Weather Rev.*, **115,** 1330–1338
Richardson, D. S.	2000	Skill and economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126,** 649–668
Roebber, P. J. and Bosart, L. F.	1996	The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting*, **11,** 544–559
Talagrand, O., Vautard, R. and Strauss, B.	1997	'Evaluation of probabilistic prediction systems'. Pp. 157–166 in Proceedings of the ECMWF workshop on predictability, 20–22 October 1997, ECMWF, Shinfield Park, Reading, UK
Wilks, D. S.	1995	*Statistical methods in the atmospheric sciences*. Academic Press
Wilks, D. S. and Hamill, T. M.	1995	Potential economic value of ensemble forecasts. *Mon. Weather Rev.*, **123,** 3565–3575